

On the consistency of a spatial-type interval-valued median for random intervals

Beatriz Sinova · Stefan Van Aelst

Received: date / Accepted: date

Abstract The sample d_θ -median is a robust estimator of the central tendency or location of an interval-valued random variable. While the interval-valued sample mean can be highly influenced by outliers, this spatial-type interval-valued median remains much more reliable. In this paper, we show that under general conditions the sample d_θ -median is a strongly consistent estimator of the d_θ -median of an interval-valued random variable.

Keywords Interval-valued data · Random intervals · Spatial median · Consistency

Mathematics Subject Classification (2000) MSC 62G35 · MSC 62–07

1 Introduction

In this data driven area, the amount and complexity of the available data grows at an almost incredible speed. Therefore, there is a high need to develop novel tools to cope with such complex data structures. Whereas the first statistical techniques were designed only to manage either quantitative or qualitative data, we can now find statistical procedures to handle functional data (see for instance Arribas-Gil and Müller 2014; Febrero-Bande and González-Manteiga 2013; Jacques and Preda 2014), fuzzy-valued data (see, for instance, Ferraro and Giordani 2013; González-Rodríguez *et al.* 2012; Coppi

B. Sinova
Departamento de Estadística e I.O. y D.M.,
Universidad de Oviedo, 33071 Oviedo, Spain
E-mail: sinovabeatriz@uniovi.es

B. Sinova · S. Van Aelst
Department of Applied Mathematics, Computer Science and Statistics,
Ghent University, 9000 Gent, Belgium

S. Van Aelst
Department of Mathematics, KU Leuven, 3001 Leuven, Belgium

et al. 2012); incomplete/missing data (see, for instance, Bianco *et al.* 2013; Ferraty *et al.* 2013; Lin 2014; Zhao *et al.* 2013), and several other types of data.

Interval-valued data are a type of complex data that requires specific statistical techniques to analyze them. Interval-valued data may arise for different reasons. In some cases the underlying random variable is intrinsically interval-valued, e.g. the daily fluctuation of the systolic blood pressure. In other cases, there is an underlying real-valued but to preserve a level of confidentiality respondents are only asked to indicate the interval containing their value, e.g. their salary. It may also happen that the real-valued measurement is only partially known due to certain limitations, such as is the case for interval censored data. Finally, aggregation of a typically large dataset may lead to e.g. interval-valued symbolic data which include interval variation and structure.

The d_θ -median considered here does not make any assumption about the source of the interval-valued data. In particular, it does not matter whether the random experiment that generated the data involves an underlying observable real-valued random variable or not. An important remark is that the space of intervals is only a semilinear space, but not a linear space due to the lack of the opposite of an interval. Therefore, although intervals can be identified with two-dimensional vectors (with first component the mid-point/centre and second component the nonnegative spread/radius), it is not advisable to treat them as regular bivariate data. Indeed, common assumptions for multivariate techniques do not hold in this case.

Statistical procedures for random interval-valued data have already been proposed in the literature for different purposes, such as regression analysis (e.g. Gil *et al.* 2002; 2007; González-Rodríguez *et al.* 2012; Blanco-Fernández *et al.* 2011; 2013; Lima Neto *et al.* 2011; Fagundes *et al.* 2013; Giordani 2014); testing hypotheses (e.g. Montenegro *et al.* 2008; Nakama *et al.* 2010; González-Rodríguez *et al.* 2012), clustering (e.g. De Carvalho *et al.* 2006; D'Urso *et al.* 2006; 2011; 2014; Giusti and Grassini 2008; Da Costa *et al.* 2013, etc.), principal component analysis (e.g. Billard and Diday 2003; D'Urso and Giordani 2004; Makosso-Kallyth and Diday 2012, etc.), modelling distributions (see Brito and Duarte Silva 2012; Sun and Ralescu 2014).

One of the most commonly used location measures is the Aumann-type mean (see Aumann 1965). It is indeed supported by numerous valuable properties, including laws of Large Numbers, and is also coherent with the interval arithmetic. The main disadvantage is that it is strongly influenced by outliers and data changes, which makes this measure not always suitable as a summary measure of the distribution of a random interval. This drawback is in fact inherited from the standard real/vectorial-valued case. In the real case, the most popular alternative is the median.

In the real case, the most popular robust alternative to the mean is the median. For multivariate data the spatial median (also called the L_1 -median, as introduced by Weber 1909) is a popular robust alternative to estimate the center of the multivariate data. The spatial median is defined as the point in multivariate space with minimal average Euclidean distance to the observations.

For more details and extensions, see for instance Gower (1974), Brown (1983), Milasevic and Ducharme (1987), (Cadre 2001), Roelant and Van Aelst (2007), Debruyne *et al.* (2010), Fritz *et al.* (2012), Zuo (2013).

Sinova and Van Aelst (2014) adapted the spatial median to interval-valued data by using a suitable L^2 metric on this space (see also Sinova *et al.* 2013). They used the versatile generalized metric introduced by Bertoluzza *et al.* (1995, see also Gil *et al.* 2002; Trutschnig *et al.* 2009) The resulting d_θ -median estimator has been shown to be robust with high breakdown point and good finite-sample properties. In this paper we show another important property of the estimator, which is its strong consistency.

The rest of this paper is organized as follows: in Section 2 the basic concepts related to the interval-valued space, interval arithmetic and metric for intervals will be introduced, as well as the usual location measure. In Section 3, the d_θ -median for random intervals and its main properties are recalled. The strong consistency of the d_θ -median is proven in Section 4. Finally, some concluding remarks are presented in Section 5.

2 The d_θ -median of a random interval

Let $\mathcal{K}_c(\mathbb{R})$ denote the class of nonempty compact intervals. Any interval K in the space $\mathcal{K}_c(\mathbb{R})$ can be characterized in terms of either its infimum and supremum, $K = [\inf K, \sup K]$, or its mid-point and spread or radius, $K = [\text{mid } K - \text{spr } K, \text{mid } K + \text{spr } K]$, where

$$\text{mid } K = \frac{\inf K + \sup K}{2}, \quad \text{spr } K = \frac{\sup K - \inf K}{2} \geq 0.$$

The usual interval arithmetic provides the addition, i.e. $K + K' = [\inf K + \inf K', \sup K + \sup K']$ with $K, K' \in \mathcal{K}_c(\mathbb{R})$ and the product by a scalar, i.e. $\gamma \cdot K = [\gamma \cdot \text{mid } K - |\gamma| \cdot \text{spr } K, \gamma \cdot \text{mid } K + |\gamma| \cdot \text{spr } K]$ with $K \in \mathcal{K}_c(\mathbb{R})$ and $\gamma \in \mathbb{R}$. With these two operations the space $\mathcal{K}_c(\mathbb{R})$ is semilinear, but not linear due to the lack of a difference of intervals. Therefore, statistical techniques for interval-valued data are based on distances.

To measure the distance between two interval-valued observations, we consider the d_θ metric introduced by Bertoluzza *et al.* (1995), which can be defined as (see Gil *et al.* 2002):

$$d_\theta(K, K') = \sqrt{(\text{mid } K - \text{mid } K')^2 + \theta \cdot (\text{spr } K - \text{spr } K')^2},$$

where $K, K' \in \mathcal{K}_c(\mathbb{R})$ and $\theta \in (0, \infty)$. Following the general random set approach, a *random interval* can usually be defined as a Borel measurable mapping $X : \Omega \rightarrow \mathcal{K}_c(\mathbb{R})$, where (Ω, \mathcal{A}, P) is a probability space with respect to \mathcal{A} and on $\mathcal{K}_c(\mathbb{R})$ the Borel σ -field generated by the topology induced by the d_θ metric. As a consequence from the Borel measurability, crucial concepts in probabilistic and inferential developments, such as the (induced) distribution of a random interval or the stochastic independence of random intervals, are well-defined.

One of the most used location measures is the *Aumann-type mean value*. It is defined, if it exists, as the interval $E[X] = [E(\inf X), E(\sup X)]$ or $E[X] = [E(\text{mid } X) - E(\text{spr } X), E(\text{mid } X) + E(\text{spr } X)]$ (both expressions are equivalent). Moreover, it is the Fréchet expectation with respect to the d_θ metric, i.e., it is the unique interval that minimizes, over $K \in \mathcal{K}_c(\mathbb{R})$, the expression $E[(d_\theta(X, K))^2]$.

As a robust alternative to the Aumann-type mean, Sinova and Van Aelst (2014) proposed the d_θ -median as measure of location, which is defined as follows.

Definition 1 The d_θ -median(s) of a random interval $X : \Omega \rightarrow \mathcal{K}_c(\mathbb{R})$ is(are) the interval(s) $M_\theta[X] \in \mathcal{K}_c(\mathbb{R})$ such that

$$E(d_\theta(X, M_\theta[X])) = \min_{K \in \mathcal{K}_c(\mathbb{R})} E(d_\theta(X, K)),$$

whenever the involved expectations exist.

Analogously, the sample d_θ -median statistic is defined as follows.

Definition 2 Let (X_1, \dots, X_n) be a simple random sample from a random interval $X : \Omega \rightarrow \mathcal{K}_c(\mathbb{R})$ with realizations $\mathbf{x}_n = (x_1, \dots, x_n)$. The *sample d_θ -median* (or medians) $\widehat{M}_\theta[X]_n$ is (are) the random interval that takes, for \mathbf{x}_n , the interval value(s) $\widehat{M}[\mathbf{x}_n]$ that is (are) the solution(s) of the following optimization problem:

$$\begin{aligned} & \min_{K \in \mathcal{K}_c(\mathbb{R})} \frac{1}{n} \sum_{i=1}^n d_\theta(x_i, K) \\ &= \min_{(y, z) \in \mathbb{R} \times [0, \infty)} \frac{1}{n} \sum_{i=1}^n \sqrt{(\text{mid } x_i - y)^2 + \theta \cdot (\text{spr } x_i - z)^2} \end{aligned}$$

where K , y and z depend on \mathbf{x}_n (which has been omitted from the notation for the sake of simplicity) and the fixed value θ .

Sinova and Van Aelst (2014) showed the existence of the sample d_θ -median estimator and its uniqueness whenever not all the two-dimensional sample points $\{(\text{mid } x_i, \text{spr } x_i)\}_{i=1}^n$ are collinear. Moreover, the robustness was shown by its finite sample breakdown point (Donoho and Huber 1983) which is given by

$$\text{fsbp}(\widehat{M}_\theta[X]_n, \mathbf{x}_n, d_\theta) = \frac{1}{n} \cdot \lfloor \frac{n+1}{2} \rfloor,$$

where $\lfloor \cdot \rfloor$ denotes the floor function.

3 Consistency of the sample d_θ -median

In this section we investigate the strong consistency of the sample d_θ -median under general conditions.

Theorem 1 *Let X be a random interval associated with a probability space (Ω, \mathcal{A}, P) such that the d_θ -median exists and is unique. Then, the sample d_θ -median is a strongly consistent estimator of the d_θ -median, that is,*

$$\lim_{n \rightarrow \infty} d_\theta(\widehat{M_\theta[X]}_n, M_\theta[X]) = 0 \quad \text{a.s.}[P].$$

Proof. Sufficient conditions for the strong consistency of an estimator are given in Huber (1967). We will check that these conditions, detailed below, are satisfied in our case:

- The parameter set $(\mathbb{R} \times [0, \infty))$ in our case, with the topology induced by the d_θ -metric) is a locally compact space with a countable base and (Ω, \mathcal{A}, P) is a probability space.

Let $\rho(\omega, (y, z))$ be the following real-valued function on $\Omega \times (\mathbb{R} \times [0, \infty))$:

$$\begin{aligned} \rho : \Omega \times (\mathbb{R} \times [0, \infty)) &\longrightarrow \mathbb{R} \\ (\omega, (y, z)) &\longmapsto d_\theta(X(\omega), [y - z, y + z]). \end{aligned}$$

- Assuming that $\omega_1, \omega_2, \dots$ are independent Ω -valued random elements with common probability distribution P , the sequence of functions $\{T_n\}_{n \in \mathbb{N}}$, defined as $T_n(\omega_1, \dots, \omega_n) = M_\theta[(X(\omega_1), \dots, X(\omega_n))]_n$, satisfies that $\frac{1}{n} \sum_{i=1}^n d_\theta(X(\omega_i), T_n(\omega_1, \dots, \omega_n)) - \inf_{(y, z) \in \mathbb{R} \times [0, \infty)} \frac{1}{n} \sum_{i=1}^n d_\theta(X(\omega_i), [y - z, y + z]) \xrightarrow[n]{} 0$ almost surely (obviously because of the definition of the sample d_θ -median).

Assumption (A-1) For each fixed $(y_0, z_0) \in \mathbb{R} \times [0, \infty)$, the function

$$\begin{aligned} \rho_0 : \Omega &\longrightarrow \mathbb{R} \\ \omega &\longmapsto \rho(\omega, (y_0, z_0)) = d_\theta(X(\omega), [y_0 - z_0, y_0 + z_0]) \\ &= \sqrt{(\text{mid } X(\omega) - y_0)^2 + \theta \cdot (\text{spr } X(\omega) - z_0)^2} \end{aligned}$$

is \mathcal{A} -measurable and separable in Doob's sense: there is a P -null set N and a countable subset $S \subset \mathbb{R} \times [0, \infty)$ such that for every open set $U \subset \mathbb{R} \times [0, \infty)$ and every closed interval A , the sets

$$\begin{aligned} V_1 &= \{\omega : \rho(\omega, (y, z)) \in A, \forall (y, z) \in U\} \\ V_2 &= \{\omega : \rho(\omega, (y, z)) \in A, \forall (y, z) \in U \cap S\} \end{aligned}$$

differ by at most a subset of N .

Assumption (A-2) The function ρ is a.s. lower semicontinuous in (y_0, z_0) , that is,

$$\inf_{(y, z) \in U} \rho(\omega, (y, z)) \longrightarrow \rho(\omega, (y_0, z_0)),$$

as the neighborhood U of (y_0, z_0) shrinks to $\{(y_0, z_0)\}$.

Assumption (A-3) There is a measurable function $a : \Omega \rightarrow \mathbb{R}$ such that

$$\begin{aligned} E[\rho(\omega, (y, z)) - a(\omega)]^- &< \infty \quad \text{for all } (y, z) \in \mathbb{R} \times [0, \infty), \\ E[\rho(\omega, (y, z)) - a(\omega)]^+ &< \infty \quad \text{for some } (y, z) \in \mathbb{R} \times [0, \infty). \end{aligned}$$

Thus, $\gamma((y, z)) = E[\rho(\omega, (y, z)) - a(\omega)]$ is well-defined for all (y, z) .

Assumption (A-4) There is a $(y_0, z_0) \in \mathbb{R} \times [0, \infty)$ such that $\gamma((y, z)) > \gamma((y_0, z_0))$ for all $(y, z) \neq (y_0, z_0)$.

Assumption (A-5) There is a continuous function $b((y, z)) > 0$ such that

- for some integrable h ,

$$\inf_{(y,z) \in \mathbb{R} \times [0, \infty)} \frac{\rho(\omega, (w, z)) - a(\omega)}{b((y, z))} \geq h(\omega).$$

- the following condition is satisfied:

$$\liminf_{(y,z) \rightarrow \infty} b((y, z)) > \gamma((y_0, z_0)).$$

- it is also fulfilled that:

$$E \left[\liminf_{(y,z) \rightarrow \infty} \frac{\rho(\omega, (y, z)) - a(\omega)}{b((y, z))} \right] \geq 1.$$

We now verify these conditions of Huber:

(A-1) For each fixed $(y_0, z_0) \in \mathbb{R} \times [0, \infty)$, the function ρ_0 is \mathcal{A} -measurable (because $\text{mid } X$ and $\text{spr } X$ are measurable functions since X is a random interval) and separable in Doob's sense: choosing $S = \mathbb{Q} \times (\mathbb{Q} \cap [0, \infty))$ as countable subset, for every open set $U \subset \mathbb{R} \times [0, \infty)$ and every closed interval A , it will be seen that the sets

$$V_1 = \{\omega : \rho_0(\omega) \in A, \forall (y, z) \in U\}, \quad V_2 = \{\omega : \rho_0(\omega) \in A, \forall (y, z) \in U \cap S\}$$

coincide. Obviously, $V_1 \subseteq V_2$. By *reductio ad absurdum*, it is now supposed that $V_2 \cap V_1^c \neq \emptyset$. Let $\omega_0 \in V_2 \cap V_1^c$:

- Since $\omega_0 \in V_2$, $\rho(\omega_0, (y, z)) \in A$ for all $(y, z) \in U \cap S$;
- Since $\omega_0 \in V_1^c$, there exists $(y_0, z_0) \in U$ such that $\rho(\omega_0, (y_0, z_0)) \in A^c$. A^c is an open set, so there exists a ball of radius $r > 0$ such that

$$(\rho(\omega_0, (y_0, z_0)) - r, \rho(\omega_0, (y_0, z_0)) + r) \subseteq A^c.$$

Notice now that, for a fixed $\omega \in \Omega$, the function

$$\begin{aligned} \rho_\omega : \mathbb{R} \times [0, \infty) &\longrightarrow \mathbb{R} \\ (y, z) &\longmapsto \rho(\omega, (y, z)) = \sqrt{(\text{mid } X(\omega) - y)^2 + \theta \cdot (\text{spr } X(\omega) - z)^2} \end{aligned}$$

is continuous. Therefore, $\rho_{\omega_0}^{-1}(\rho(\omega_0, (y_0, z_0)) - r, \rho(\omega_0, (y_0, z_0)) + r)$ is an open set of $\mathbb{R} \times [0, \infty)$ and $U \cap \rho_{\omega_0}^{-1}(\rho(\omega_0, (y_0, z_0)) - r, \rho(\omega_0, (y_0, z_0)) + r) \neq \emptyset$ too. S is a dense set of $\mathbb{R} \times [0, \infty)$, so

$$U \cap \rho_{\omega_0}^{-1}(\rho(\omega_0, (y_0, z_0)) - r, \rho(\omega_0, (y_0, z_0)) + r) \cap S \neq \emptyset.$$

Let $(y', z') \in U \cap \rho_{\omega_0}^{-1}(\rho(\omega_0, (y_0, z_0)) - r, \rho(\omega_0, (y_0, z_0)) + r) \cap S$. Then, $(y', z') \in U \cap S$, so $\rho(\omega_0, (y', z')) \in A$. But also,

$$\rho(\omega_0, (y', z')) \in (\rho(\omega_0, (y_0, z_0)) - r, \rho(\omega_0, (y_0, z_0)) + r) \subset A^c.$$

This is a contradiction, so the conclusion is that $V_2 \subseteq V_1$.

(A-2) Indeed, it will be proved for all $\omega \in \Omega$. Let ω be any element of Ω and let (y_0, z_0) be any (fixed) point of $\mathbb{R} \times [0, \infty)$.

First, notice that it is fulfilled for a sequence of neighborhoods $\{U_n\}_{n \in \mathbb{N}}$ of (y_0, z_0) when $U_n \supseteq U_{n+1}$ for all n that

$$\left\{ \inf_{(y,z) \in U_n} d_\theta(X(\omega), [y-z, y+z]) \right\}_{n \in \mathbb{N}}$$

is a monotonically increasing sequence. Furthermore, this sequence is bounded since

$$\inf_{(y,z) \in U_n} d_\theta(X(\omega), [y-z, y+z]) \leq d_\theta(X(\omega), [y_0-z_0, y_0+z_0])$$

for all $n \in \mathbb{N}$ because $(y_0, z_0) \in \cap_{n \in \mathbb{N}} U_n$. Therefore, the sequence converges to its supremum, which will be $d_\theta(X(\omega), [y_0-z_0, y_0+z_0])$.

By *reductio ad absurdum*, suppose that there is a smaller upper bound

$$c = d_\theta(X(\omega), [y_0-z_0, y_0+z_0]) - \varepsilon,$$

for an arbitrary $\varepsilon > 0$. Let's denote by U_{n_0} a neighborhood of (y_0, z_0) satisfying that $U_{n_0} \subseteq B((y_0, z_0), \frac{\varepsilon}{2})$. Then, it can be seen that

$$c < \inf_{(y,z) \in U_{n_0}} d_\theta(X(\omega), [y-z, y+z]),$$

so c cannot be the supremum. Thus, using the triangular inequality,

$$\begin{aligned} & \inf_{(y,z) \in U_{n_0}} d_\theta(X(\omega), [y-z, y+z]) \geq \inf_{(y,z) \in B((y_0, z_0), \frac{\varepsilon}{2})} d_\theta(X(\omega), [y-z, y+z]) \\ & \geq \inf_{(y,z) \in B((y_0, z_0), \frac{\varepsilon}{2})} [d_\theta(X(\omega), [y_0-z_0, y_0+z_0]) - d_\theta([y-z, y+z], [y_0-z_0, y_0+z_0])] \\ & = d_\theta(X(\omega), [y_0-z_0, y_0+z_0]) - \sup_{(y,z) \in B((y_0, z_0), \frac{\varepsilon}{2})} d_\theta([y-z, y+z], [y_0-z_0, y_0+z_0]) \\ & > d_\theta(X(\omega), [y_0-z_0, y_0+z_0]) - \varepsilon = c. \end{aligned}$$

Now this result will be extended to general sequences $\{U_n\}_{n \in \mathbb{N}}$. Consider the suprema and the infima radii reached in every neighborhood, namely,

$$\begin{aligned} r_n &= \sup_{(y,z) \in U_n} d_\theta([y_0-z_0, y_0+z_0], [y-z, y+z]), \\ s_n &= \inf_{(y,z) \in U_n} d_\theta([y_0-z_0, y_0+z_0], [y-z, y+z]). \end{aligned}$$

It is known that $r_n \xrightarrow{n} 0$, since $\{U_n\}_{n \in \mathbb{N}}$ shrinks to $\{(y_0, z_0)\}$. Moreover, $s_n \xrightarrow{n} 0$ as $0 \leq s_n \leq r_n$ for all $n \in \mathbb{N}$.

Let ε be any nonnegative number. As $r_n \xrightarrow{n} 0$, there exists $n_1 \in \mathbb{N}$ such that for all $n > n_1$, $r_n < \varepsilon$. Then, $U_n \subseteq B((y_0, z_0), r_n)$ and

$$\begin{aligned} & \inf_{(y,z) \in U_n} d_\theta(X(\omega), [y-z, y+z]) \geq \inf_{(y,z) \in B((y_0, z_0), r_n)} d_\theta(X(\omega), [y-z, y+z]) \\ & \geq d_\theta(X(\omega), [y_0-z_0, y_0+z_0]) - \sup_{(y,z) \in B((y_0, z_0), r_n)} d_\theta([y_0-z_0, y_0+z_0], [y-z, y+z]) \end{aligned}$$

$$> d_\theta(X(\omega), [y_0 - z_0, y_0 + z_0]) - \varepsilon.$$

Analogously, as $s_n \xrightarrow{n} 0$, there exists $n_2 \in \mathbb{N}$ such that for all $n > n_2$, $s_n < \varepsilon$. Therefore, $U_n \supseteq B((y_0, z_0), s_n)$ and

$$\begin{aligned} \inf_{(y,z) \in U_n} d_\theta(X(\omega), [y - z, y + z]) &\leq \inf_{(y,z) \in B((y_0, z_0), s_n)} d_\theta(X(\omega), [y - z, y + z]) \\ &\leq d_\theta(X(\omega), [y_0 - z_0, y_0 + z_0]) + \inf_{(y,z) \in B((y_0, z_0), s_n)} d_\theta([y - z, y + z], [y_0 - z_0, y_0 + z_0]) \\ &< d_\theta(X(\omega), [y_0 - z_0, y_0 + z_0]) + \varepsilon. \end{aligned}$$

So for any $\varepsilon > 0$, there exists $n_0 = \max\{n_1, n_2\}$, such that for all $n > n_0$,

$$\begin{aligned} d_\theta(X(\omega), [y_0 - z_0, y_0 + z_0]) - \varepsilon &< \inf_{(y,z) \in U_n} d_\theta(X(\omega), [y - z, y + z]) \\ &< d_\theta(X(\omega), [y_0 - z_0, y_0 + z_0]) + \varepsilon, \end{aligned}$$

that is to say,

$$\left| \inf_{(y,z) \in U_n} d_\theta(X(\omega), [y - z, y + z]) - d_\theta(X(\omega), [y_0 - z_0, y_0 + z_0]) \right| < \varepsilon,$$

so the sequence $\left\{ \inf_{(y,z) \in U_n} d_\theta(X(\omega), [y - z, y + z]) \right\}_{n \in \mathbb{N}}$ converges to $d_\theta(X(\omega), [y_0 - z_0, y_0 + z_0])$.

(A-3) Let a be the measurable function (see (A-1)):

$$\begin{aligned} a : \Omega &\longrightarrow \mathbb{R} \\ \omega &\longmapsto d_\theta(X(\omega), [0, 0]) = \sqrt{(\text{mid } X(\omega))^2 + \theta \cdot (\text{spr } X(\omega))^2}. \end{aligned}$$

Fixed any arbitrary $(y, z) \in \mathbb{R} \times [0, \infty)$,

$$\begin{aligned} &E[\rho(\omega, (y, z)) - a(\omega)]^- \\ &= \int_{\Omega} -\min\{d_\theta(X(\omega), [y - z, y + z]) - d_\theta(X(\omega), [0, 0]), 0\} dP(\omega) \\ &= \int_{\{\omega \in \Omega : d_\theta(X(\omega), [0, 0]) > d_\theta(X(\omega), [y - z, y + z])\}} [d_\theta(X(\omega), [0, 0]) - d_\theta(X(\omega), [y - z, y + z])] dP(\omega). \end{aligned}$$

By the triangular inequality,

$$\begin{aligned} &\leq \int_{\{\omega \in \Omega : d_\theta(X(\omega), [0, 0]) > d_\theta(X(\omega), [y - z, y + z])\}} [d_\theta(X(\omega), [y - z, y + z]) + d_\theta([y - z, y + z], [0, 0]) \\ &\quad - d_\theta(X(\omega), [y - z, y + z])] dP(\omega) \\ &= d_\theta([y - z, y + z], [0, 0]) \cdot P(\omega : d_\theta(X(\omega), [0, 0]) > d_\theta(X(\omega), [y - z, y + z])) < \infty. \end{aligned}$$

Analogously,

$$E[\rho(\omega, (y, z)) - a(\omega)]^+$$

$$\begin{aligned}
&= \int_{\Omega} \max\{d_{\theta}(X(\omega), [y-z, y+z]) - d_{\theta}(X(\omega), [0, 0]), 0\} dP(\omega) \\
&= \int_{\{\omega \in \Omega : d_{\theta}(X(\omega), [0, 0]) \leq d_{\theta}(X(\omega), [y-z, y+z])\}} [d_{\theta}(X(\omega), [y-z, y+z]) - d_{\theta}(X(\omega), [0, 0])] dP(\omega).
\end{aligned}$$

By the triangular inequality,

$$\begin{aligned}
&\leq \int_{\{\omega \in \Omega : d_{\theta}(X(\omega), [0, 0]) \leq d_{\theta}(X(\omega), [y-z, y+z])\}} [d_{\theta}(X(\omega), [0, 0]) + d_{\theta}([0, 0], [y-z, y+z]) \\
&\quad - d_{\theta}(X(\omega), [0, 0])] dP(\omega) \\
&= d_{\theta}([0, 0], [y-z, y+z]) \cdot P(\omega : d_{\theta}(X(\omega), [0, 0]) \leq d_{\theta}(X(\omega), [y-z, y+z])) < \infty.
\end{aligned}$$

So the second inequality also holds for all $(y, z) \in \mathbb{R} \times [0, \infty)$ in this case.

(A-4) The d_{θ} -median exists and is unique, so that

$$\begin{aligned}
(\text{mid } M_{\theta}[X], \text{spr } M_{\theta}[X]) &= \arg \min_{(y, z) \in \mathbb{R} \times [0, \infty)} E[d_{\theta}(X(\omega), [y-z, y+z])] \\
&= \arg \min_{(y, z) \in \mathbb{R} \times [0, \infty)} E[d_{\theta}(X(\omega), [y-z, y+z])] - E[d_{\theta}(X(\omega), [0, 0])] \\
&= \arg \min_{(y, z) \in \mathbb{R} \times [0, \infty)} E[d_{\theta}(X(\omega), [y-z, y+z]) - d_{\theta}(X(\omega), [0, 0])] \\
&= \arg \min_{(y, z) \in \mathbb{R} \times [0, \infty)} \gamma((y, z))
\end{aligned}$$

and $(y_0, z_0) := (\text{mid } M_{\theta}[X], \text{spr } M_{\theta}[X])$ fulfills this assumption.

(A-5) There is a continuous function $b((y, z)) > 0$

$$\begin{aligned}
b : \mathbb{R} \times [0, \infty) &\longrightarrow \mathbb{R} \\
(y, z) &\longmapsto d_{\theta}([y-z, y+z], [0, 0]) + 1
\end{aligned}$$

such that

– for the integrable function $h(\omega) := -1$,

$$\inf_{(y, z) \in \mathbb{R} \times [0, \infty)} \frac{d_{\theta}(X(\omega), [y-z, y+z]) - d_{\theta}(X(\omega), [0, 0])}{d_{\theta}([y-z, y+z], [0, 0]) + 1} \geq -1$$

because using the triangular inequality,

$$\begin{aligned}
&\inf_{(y, z) \in \mathbb{R} \times [0, \infty)} \frac{d_{\theta}(X(\omega), [y-z, y+z]) - d_{\theta}(X(\omega), [0, 0])}{d_{\theta}([y-z, y+z], [0, 0]) + 1} \\
&\geq \inf_{(y, z) \in \mathbb{R} \times [0, \infty)} \frac{d_{\theta}(X(\omega), [0, 0]) - d_{\theta}([y-z, y+z], [0, 0]) - d_{\theta}(X(\omega), [0, 0])}{d_{\theta}([y-z, y+z], [0, 0]) + 1} \\
&= \inf_{(y, z) \in \mathbb{R} \times [0, \infty)} \frac{-d_{\theta}([y-z, y+z], [0, 0])}{d_{\theta}([y-z, y+z], [0, 0]) + 1} \geq -1.
\end{aligned}$$

– the following condition is satisfied:

$$\liminf_{(y,z) \rightarrow \infty} b((y, z)) > \gamma((y_0, z_0)).$$

Let $\{(y_n, z_n)\} \subset \mathbb{R} \times [0, \infty)$ be any sequence with $(y_n, z_n) \xrightarrow[n]{} \infty$ (i.e., $d_\theta([y_n - z_n, y_n + z_n], [0, 0]) \xrightarrow[n]{} \infty$) and

$$M = E[d_\theta(X(\omega), [y_0 - z_0, y_0 + z_0]) - d_\theta(X(\omega), [0, 0])] = \gamma((y_0, z_0)) \in \mathbb{R},$$

where (y_0, z_0) represents the minimum found in (A-4). Then, there exists $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$,

$$d_\theta([y_n - z_n, y_n + z_n], [0, 0]) > M.$$

So, for all $n \geq n_0$,

$$\inf_{k \geq n} b((y_k, z_k)) = \inf_{k \geq n} (d_\theta([y_k - z_k, y_k + z_k], [0, 0]) + 1) \geq M + 1.$$

Finally,

$$\liminf_{n \rightarrow \infty} b((y_n, z_n)) = \lim_{n \rightarrow \infty} (\inf_{k \geq n} b((y_k, z_k))) \geq M + 1 > M = \gamma((y_0, z_0)).$$

– it is also fulfilled that:

$$E \left[\liminf_{(y,z) \rightarrow \infty} \frac{d_\theta(X(\omega), [y - z, y + z]) - d_\theta(X(\omega), [0, 0])}{b((y, z))} \right] \geq 1.$$

Let's see that

$$\liminf_{(y,z) \rightarrow \infty} \frac{d_\theta(X(\omega), [y - z, y + z]) - d_\theta(X(\omega), [0, 0])}{d_\theta([y - z, y + z], [0, 0]) + 1} \geq 1,$$

so the result follows.

$$\begin{aligned} & \liminf_{(y,z) \rightarrow \infty} \frac{d_\theta(X(\omega), [y - z, y + z]) - d_\theta(X(\omega), [0, 0])}{d_\theta([y - z, y + z], [0, 0]) + 1} \\ &= \lim_{n \rightarrow \infty} \left(\inf_{k \geq n} \frac{d_\theta(X(\omega), [y_k - z_k, y_k + z_k]) - d_\theta(X(\omega), [0, 0])}{d_\theta([y_k - z_k, y_k + z_k], [0, 0]) + 1} \right) \end{aligned}$$

for any fixed $\omega \in \Omega$. The sequence

$$\left\{ \inf_{k \geq n} \frac{d_\theta(X(\omega), [y_k - z_k, y_k + z_k]) - d_\theta(X(\omega), [0, 0])}{d_\theta([y_k - z_k, y_k + z_k], [0, 0]) + 1} \right\}_{n \in \mathbb{N}}$$

is monotonically increasing and is upper bounded by 1: for all $k \in \mathbb{N}$, using the triangular inequality,

$$\frac{d_\theta(X(\omega), [y_k - z_k, y_k + z_k]) - d_\theta(X(\omega), [0, 0])}{d_\theta([y_k - z_k, y_k + z_k], [0, 0]) + 1}$$

$$\leq \frac{d_\theta([y_k - z_k, y_k + z_k], [0, 0])}{d_\theta([y_k - z_k, y_k + z_k], [0, 0]) + 1} \leq 1.$$

So it converges to its supremum:

$$\begin{aligned} & \lim_{n \rightarrow \infty} \left(\inf_{k \geq n} \frac{d_\theta(X(\omega), [y_k - z_k, y_k + z_k]) - d_\theta(X(\omega), [0, 0])}{d_\theta([y_k - z_k, y_k + z_k], [0, 0]) + 1} \right) \\ &= \sup_n \left(\inf_{k \geq n} \frac{d_\theta(X(\omega), [y_k - z_k, y_k + z_k]) - d_\theta(X(\omega), [0, 0])}{d_\theta([y_k - z_k, y_k + z_k], [0, 0]) + 1} \right) \end{aligned}$$

Let's finally see that this supremum is at least equal to 1. By *reductio ad absurdum*, let's suppose that

$$\sup_n \left(\inf_{k \geq n} \frac{d_\theta(X(\omega), [y_k - z_k, y_k + z_k]) - d_\theta(X(\omega), [0, 0])}{d_\theta([y_k - z_k, y_k + z_k], [0, 0]) + 1} \right) = 1 - \varepsilon,$$

for some $\varepsilon > 0$. One gets then a contradiction because one finds an $n^* \in \mathbb{N}$ such that

$$\inf_{k \geq n^*} \frac{d_\theta(X(\omega), [y_k - z_k, y_k + z_k]) - d_\theta(X(\omega), [0, 0])}{d_\theta([y_k - z_k, y_k + z_k], [0, 0]) + 1} > 1 - \varepsilon$$

since for all $k \geq n^*$,

$$\frac{d_\theta(X(\omega), [y_k - z_k, y_k + z_k]) - d_\theta(X(\omega), [0, 0])}{d_\theta([y_k - z_k, y_k + z_k], [0, 0]) + 1} \geq 1 - \frac{\varepsilon}{2} > 1 - \varepsilon$$

as we will show now. Recall that $(y_n, z_n) \xrightarrow[n]{n} \infty$, so for all $M \in \mathbb{R}$, there exists $n^* \in \mathbb{N}$ such that for all $n \geq n^*$, $d_\theta([y_n - z_n, y_n + z_n], [0, 0]) > M$. Therefore,

$$\begin{aligned} d_\theta([y_n - z_n, y_n + z_n], X(\omega)) &\geq d_\theta([y_n - z_n, y_n + z_n], [0, 0]) - d_\theta(X(\omega), [0, 0]) \\ &> M - d_\theta(X(\omega), [0, 0]). \end{aligned}$$

Taking $M := \frac{2}{\varepsilon} - 1 + \frac{4}{\varepsilon} \cdot d_\theta(X(\omega), [0, 0]) \in \mathbb{R}$ (for the fixed arbitrary $\omega \in \Omega$), we can easily check that $1 - \frac{\varepsilon}{2}$ is a lower bound of the sequence

$$\left\{ \frac{d_\theta(X(\omega), [y_k - z_k, y_k + z_k]) - d_\theta(X(\omega), [0, 0])}{d_\theta([y_k - z_k, y_k + z_k], [0, 0]) + 1} \right\}_{k \geq n^*}.$$

For any $k \geq n^*$,

$$\begin{aligned} & d_\theta(X(\omega), [y_k - z_k, y_k + z_k]) - d_\theta(X(\omega), [0, 0]) \\ &= \left(1 - \frac{\varepsilon}{2}\right) d_\theta(X(\omega), [y_k - z_k, y_k + z_k]) + \frac{\varepsilon}{2} d_\theta(X(\omega), [y_k - z_k, y_k + z_k]) \\ &\quad - d_\theta(X(\omega), [0, 0]) \\ &\geq \left(1 - \frac{\varepsilon}{2}\right) d_\theta([y_k - z_k, y_k + z_k], [0, 0]) - \left(1 - \frac{\varepsilon}{2}\right) d_\theta(X(\omega), [0, 0]) \\ &\quad + \frac{\varepsilon}{2} d_\theta(X(\omega), [y_k - z_k, y_k + z_k]) - d_\theta(X(\omega), [0, 0]) \end{aligned}$$

$$\begin{aligned}
&= \left(1 - \frac{\varepsilon}{2}\right) d_{\theta}([y_k - z_k, y_k + z_k], [0, 0]) + \frac{\varepsilon}{2} d_{\theta}(X(\omega), [y_k - z_k, y_k + z_k]) \\
&\quad - \left(2 - \frac{\varepsilon}{2}\right) d_{\theta}(X(\omega), [0, 0]) \\
&> \left(1 - \frac{\varepsilon}{2}\right) d_{\theta}([y_k - z_k, y_k + z_k], [0, 0]) + \frac{\varepsilon}{2} \left(\frac{2}{\varepsilon} - 1 + \left(\frac{4}{\varepsilon} - 1\right) d_{\theta}(X(\omega), [0, 0])\right) \\
&\quad - \left(2 - \frac{\varepsilon}{2}\right) d_{\theta}(X(\omega), [0, 0]) = \left(1 - \frac{\varepsilon}{2}\right) d_{\theta}([y_k - z_k, y_k + z_k], [0, 0]) + 1 - \frac{\varepsilon}{2} \\
&\quad = \left(1 - \frac{\varepsilon}{2}\right) (d_{\theta}([y_k - z_k, y_k + z_k], [0, 0]) + 1). \quad \square
\end{aligned}$$

4 Concluding remarks

This paper complements the study of the properties of the d_{θ} -median as a robust estimator of the center of a random interval by showing its strong consistency which is one of the most important basic properties of an estimator. We obtained this result by showing that all the sufficient conditions of Huber (1967) are fulfilled. These results open the door to further develop robust statistical inference for random intervals based on the d_{θ} -median such as the development of hypotheses testing procedures.

Acknowledgements Authors are grateful to María Ángeles Gil for her helpful suggestions to improve this paper. The research by Beatriz Sinova was partially supported by/benefited from the Spanish Ministry of Science and Innovation Grant MTM2009-09440-C02-01. She has been also granted with the Ayuda del Programa de FPU AP2009-1197 from the Spanish Ministry of Education and the Ayuda para Estancias Breves del Programa FPU EST12/00344, an Ayuda de Investigación 2011 from the Fundación Banco Herrero and three Short Term Scientific Missions associated with the COST Action IC0702. The research by Stefan Van Aelst was supported by a grant of the Fund for Scientific Research-Flanders (FWO-Vlaanderen) and by IAP research network grant nr. P7/06 of the Belgian government (Belgian Science Policy). Their financial support is gratefully acknowledged.

References

- Arribas-Gil A, Müller H-G (2014) Pairwise dynamic time warping for event data. *Comput Stat Data Anal* 69:255–268
- Aumann RJ (1965) Integrals of set-valued functions. *J Math Anal Appl* 12:1–12
- Bertoluzza C, Corral N, Salas A (1995) On a new class of distances between fuzzy numbers. *Math & Soft Comput* 2:71–84
- Bianco AM, Boente G, Rodrigues IM (2013) Robust tests in generalized linear models with missing responses. *Comput Stat Data Anal* 65:80–97
- Billard L, Diday E (2003) From the Statistics of data to the Statistics of knowledge: Symbolic Data Analysis. *J Am Stat Ass* 98:470–487
- Blanco-Fernández A, Corral N, González-Rodríguez G (2011) Estimation of a flexible simple linear model for interval data based on set arithmetic. *Comput Stat Data Anal* 55(9):2568–2578
- Blanco-Fernández A, Colubi A, García-Bárcana M (2013) A set arithmetic-based linear regression model for modelling interval-valued responses through real-valued variables. *Inform Sci* 247:109–122

- Brito P, Duarte Silva AP (2012) Modelling interval data with Normal and Skew-Normal distributions. *J Appl Stat* 39(1):3–20
- Brown BM (1983) Statistical uses of the spatial median. *J Royal Stat Soc Ser B* 45(1):25–30
- Cadre B (2001) Convergent estimators for the L_1 -median of a Banach valued random variable. *Statistics* 35:509–521
- De Carvalho FDAT, Brito P, Bock HH (2006) Dynamic clustering for interval data based on L_2 distance. *Comput Stat* 21(2):231–250
- Coppi R, D’Urso P, Giordani P (2012) Fuzzy and possibilistic clustering for fuzzy data. *Comput Stat Data Anal* 56(4):915–927
- Da Costa AFBF, Pimentel BA, De Souza RMCR (2013) Clustering interval data through kernel-induced feature space. *J Intell Inf Syst* 40(1):101–140
- Debruyne M, Hubert M, Van Horebeek J (2010) Detecting influential observations in Kernel PCA. *Comput Stat Data Anal* 54(12):3007–3019
- Donoho DL, Huber PJ (1983) The notion of breakdown point. In: Bickel PJ, Doksum K, Hodges Jr JL (Eds.) *A Festschrift for Erich L. Lehmann*. Wadsworth, Belmont.
- Milasevic P, Ducharme GR (1987) Uniqueness of the spatial median. *Ann Statist* 15:1332–1333
- D’Urso P, Giordani P (2004) A least squares approach to principal component analysis for interval valued data. *Chemometr Intell Lab* 70(2):179–192
- D’Urso P, Giordani P (2006) A robust fuzzy k-means clustering model for interval valued data. *Comput Stat* 21(2):251–269
- D’Urso P, De Giovanni L (2011) Midpoint radius self-organizing maps for interval-valued data with telecommunications application. *Appl Soft Comput* 11(5):3877–3886
- D’Urso P, De Giovanni L, Massari R (2014) Trimmed fuzzy clustering for interval-valued data. *Adv Data Anal Classif*, in press (doi:10.1007/s11634-014-0169-3)
- Fagundes RAA, De Souza RMCR, Cysneiros FJA (2013) Robust regression with application to symbolic interval data. *Eng Appl Art Intel* 26(1):564–573
- Febrero-Bande M, González-Manteiga W (2013) Generalized additive models for functional data. *Test* 22(2):278–292
- Ferraro MB, Giordani P (2013) On possibilistic clustering with repulsion constraints for imprecise data. *Inform Sci* 245:63–75
- Ferraty F, Sued M, Vieu P (2013) Mean estimation with data missing at random for functional covariables. *Statistics* 47(4):688–706
- Fritz H, Filzmoser P, Croux C (2012) A comparison of algorithms for the multivariate L_1 -median. *Comput Stat* 27(3):393–410
- Gil MA, Lubiano MA, Montenegro M, López-García MT (2002) Least squares fitting of an affine function and strength of association for interval data. *Metrika* 56:97–111
- Gil MA, González-Rodríguez G, Colubi A, Montenegro M (2007) Testing linear independence in linear models with interval-valued data. *Comput Stat Data Anal* 51(6):3002–3015
- Giordani P (2014) Lasso-constrained regression analysis for interval-valued data. *Adv Data Anal Classif*, in press (doi:10.1007/s11634-014-0164-8)
- Giusti A, Grassini L (2008) Cluster analysis of census data using the symbolic data approach. *Adv Data Anal Classif* 2(2):163–176
- González-Rodríguez G, Blanco A, Corral N, Colubi A (2007) Least squares estimation of linear regression models for convex compact random sets. *Adv Data Anal Classif* 1(1):67–81
- González-Rodríguez G, Colubi A, Gil MA (2012) Fuzzy data treated as functional data: A one-way ANOVA test approach. *Comput Stat Data Anal* 56:943–955
- Gower JC (1974) Algorithm AS 78: The mediancentre. *Appl Statist* 23:466–470
- Huber PJ (1967) The behavior of maximum likelihood estimates under nonstandard conditions. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 1, pp. 221–233
- Ishibuchi H, Tanaka H (1990) Multiobjective programming in optimization of the interval objective function. *Europ J Oper Res* 48:219–225
- Jacques J, Preda C (2014) Model-based clustering for multivariate functional data. *Comput Stat Data Anal* 71:92–106
- Lima Neto EA, Cordeiro GM, de Carvalho FAT (2011) Bivariate symbolic regression models for interval-valued variables. *J Stat Comput Simul* 81(11):1727–1744

- Lin TH (2014) Model selection information criteria in latent class models with missing data and contingency question. *J Stat Comput Simul* 84(1):159–170
- Liu RY (1990). On a notion of data depth based on random simplices. *Ann Statist* 18:405–414
- Makosso-Kallyth S, Diday E (2012) Adaptation of interval PCA to symbolic histogram variables. *Adv Data Anal Classif* 6(2):147–159
- Montenegro M, Casals MR, Colubi A, Gil MA (2008) Testing ‘two-sided’ hypothesis about the mean of an interval-valued random set. In: Dubois D, Lubiano MA, Prade H, Gil MA, Grzegorzewski P, Hryniewicz O (Eds.) *Soft Methods for Handling Variability and Imprecision*. Springer, Heidelberg, pp. 133–139
- Nakama T, Colubi A, Lubiano MA (2010) Two-way analysis of variance for interval-valued data. In: Borgelt C, González-Rodríguez G, Trutschnig W, Lubiano MA, Gil MA, Grzegorzewski P, Hryniewicz O (Eds.) *Combining Soft Computing and Statistical Methods in Data Analysis*. Springer, Heidelberg, pp. 475–482
- Roelant E, Van Aelst S (2007) An L1-type estimator of multivariate location. *Stat Meth & Appl* 15:381–393
- Rousseeuw PJ, Ruts I (1996) Bivariate location depth. *J. Roy. Statist. Soc. Ser. C* 45:516–526
- Rousseeuw PJ, Ruts I (1998) Constructing the bivariate Tukey median. *Statistica Sinica* 8:827–839
- Sinova B, Gil MA, Colubi A, Van Aelst S (2012) The median of a random fuzzy number. The 1-norm distance approach. *Fuzzy Sets Syst* 200:99–115
- Sinova B, González-Rodríguez G, Van Aelst S (2013) An alternative approach to the median of a random interval using an L^2 metric. In: Kruse R, Berthold MR, Moewes C, Gil MA, Grzegorzewski P, Hryniewicz O (Eds.) *Sinergies of Soft Computing and Statistics for Intelligent Data Analysis*. Springer, Heidelberg, pp. 273–281
- Sun Y, Ralescu DA (2014) A normal hierarchical model and minimum contrast estimation for random intervals. *Ann Inst Stat Math*, in press (doi:10.1007/s10463-014-0453-1)
- Trutschnig W, González-Rodríguez G, Colubi A, Gil MA (2009) A new family of metrics for compact, convex (fuzzy) sets based on a generalized concept of mids and spread. *Inf Sci* 179:3964–3972
- Tukey JW (1975) Mathematics and the picturing of data. In: *Proc. International Congress of Mathematicians, Vancouver, 1994*, 2, pp. 523–531
- Vitale RA (1985) L_p metrics for compact, convex sets. *J Approx Theory* 45:280–287
- Weber A (1909) *Über den Standort der Industrien*. Mohr, Tübingen.
- Zhao P-Y, Tang M-L, Tang N-S (2013) Robust estimation of distribution functions and quantiles with non-ignorable missing data. *Canad J Stat* 41(4):575–595
- Zuo Y (2013) Multidimensional medians and uniqueness. *Comput Stat Data Anal* 66:82–88